Experience Evaluations for Human-Computer Co-Creative Processes : Planning and Conducting an Evaluation in Practice

Kantosalo, Anna

2019-01-02

Kantosalo , A & Riihiaho , S 2019 , ' Experience Evaluations for Human-Computer Co-Creative Processes : Planning and Conducting an Evaluation in Practice ' , Connection Science , vol. 31 , no. 1 , pp. 60-81 . https://doi.org/10.1080/09540091.2018.1432566

http://hdl.handle.net/10138/300089
https://doi.org/10.1080/09540091.2018.1432566

Taylor & Francis
Taylor & Francis Group

Connection Science

ISSN: 0954-0091 (Print) 1360-0494 (Online) Journal homepage: https://www.tandfonline.com/loi/ccos20

Experience evaluations for human–computer co-creative processes – planning and conducting an evaluation in practice

Anna Kantosalo & Sirpa Riihiaho

To cite this article: Anna Kantosalo & Sirpa Riihiaho (2019) Experience evaluations for human–computer co-creative processes – planning and conducting an evaluation in practice, Connection Science, 31:1, 60-81, DOI: 10.1080/09540091.2018.1432566

To link to this article: https://doi.org/10.1080/09540091.2018.1432566

Published online: 24 Aug 2018.

CONNECTION SCIENCE
2019, VOL. 31, NO. 1, 60–81
https://doi.org/10.1080/09540091.2018.1432566

Experience evaluations for human–computer co-creative processes – planning and conducting an evaluation in practice

Anna Kantosalo[a] and Sirpa Riihiaho[b]

[a]Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, Finland; [b]Department of Computer Science, University of Helsinki, Helsinki, Finland

ABSTRACT

In human–computer co-creativity, humans and creative computational algorithms create together. Too often, only the creative algorithms and their outcomes are evaluated when studying these co-creative processes, leaving the human participants to little attention. This paper presents a case study emphasising the human experiences when evaluating the use of a co-creative poetry writing system called the Poetry Machine. The co-creative process was evaluated using seven metrics: Fun, Enjoyment, Expressiveness, Outcome satisfaction, Collaboration, Ease of writing, and Ownership. The metrics were studied in a comparative setting using three co-creation processes: a human–computer, a human–human, and a human–human–computer co-creation process. Twelve pupils of age 10–11 attended the studies in six pairs trying out all the alternative writing processes. The study methods included observation in paired-user testing, questionnaires, and interview. The observations were complemented with analyses of the video recordings of the evaluation sessions. According to statistical analyses, Collaboration was the strongest in human–human–computer co-creation, and weakest in human–computer co-creation. Ownership was just the opposite: weakest in human–human–computer co-creation, and strongest in human–computer co-creation. Other metrics did not produce statistically significant results. In addition to the results, this paper presents the lessons learned in the evaluations with children using the selected methods.

ARTICLE HISTORY
Received 29 July 2017
Accepted 11 January 2018

KEYWORDS
Computational creativity; human–computer co-creativity; user experience; evaluation metrics; child–computer interaction

1. Introduction

This paper proposes a human perspective in evaluation of human–computer co-creative processes in which humans and computationally creative systems create together. Computational creativity is a sub-field of artificial intelligence devoted to the research and simulation of creative behaviour. An important goal of computational creativity is to generate creative artefacts via computational means. Human–computer co-creativity examines how we can use these computational creativity methods to promote human creativity and *vice versa*. In the human–computer co-creative process, the human becomes an integral part of the creative process itself instead of being just a part of the audience. This

CONTACT Anna Kantosalo anna.kantosalo@helsinki.fi

requires adopting a new evaluation stance, which goes beyond the traditional computational creativity evaluation foci of creative output and internal workings of the system, and includes the experiences of the human working with the system. Investigating the real user experiences of real users is essential for further developing these systems.

User experience is closely related to the domain of Interaction Design, which has been suggested as a useful evaluation paradigm for computational creativity by Bown (2014), and successfully utilised in some user evaluations of computational creativity applications (see e.g. Kantosalo, Toivanen, & Toivonen, 2015; Bown, 2015; Jacob & Magerko, 2015; Davis, Hsiao, Yashraj Singh, Li, & Magerko, 2016). The term user experience refers to a number of qualitative and hedonistic aspects of interaction, including for example "the result of enjoyable interactions and/or anticipated interactions with a product" (Lachner, Naegelein, Kowalski, Spann, & Butz, 2016). The experience oriented perspective thereby takes the complex emotive expectations and reactions to computationally creative software into account. These states and actions inevitably affect the creative use of human–computer co-creative systems.

So far, a focus on user experience has been scarce in computational creativity evaluation, but Yee-King and d'Inverno (2016) suggest that measuring the experiences of a human interacting with a system is more compelling than measuring the creativity of the system itself, when designing new co-creative systems. Similarly, Bown (2015) considers that instead of measuring the creativity of an interactive system, the complex network of interactions surrounding it should be in focus.

This study compares three co-creative processes: A human–computer (H–C), a human–human–computer (H–H–C), and a human–human (H–H) co-creative process. The computational participant of the co-creative processes is a poetry writing system called the Poetry Machine. The Poetry Machine is designed to collaborate with children at school and it is intended to help pupils create their own poetry in and out of the classroom.

This paper outlines an approach to investigating user experience in a comparative setting, comparing alternative co-creative processes within one domain. The evaluation uses seven metrics: Fun, Enjoyment, Expressiveness, Outcome Satisfaction, Ease of Writing, Collaboration, and Ownership. The metrics originate from three perspectives: user experience evaluation, computational creativity evaluation and creativity support system evaluation. The metrics are evaluated with twelve 10–11-year-old pupils in a school context, using observations in paired-user testing, questionnaires, and interview. The observations are complemented with analysis of video records of the evaluation sessions. This paper focuses on the statistical results of the case study and discusses the usefulness of and practical experiences with the chosen metrics and methodology within the study context.

2. Study context

The computational creativity system used in the study is an artificial intelligence-based system called the Poetry Machine. It is developed by a multidisciplinary team of computer scientists and pedagogical researchers. The system is based on algorithms capable of producing poetry autonomously. Cherry and Latulipe (2014) define creativity support systems as any tool or combination of tools that can help "in the open-ended creation of new artefacts". Therefore, the Poetry Machine can be considered as a creativity support system. However, the autonomous capability of the system gives the Poetry Machine more creative

Figure 1. The Poetry Machine writing interface, showing one of the poems written in this experiment.

ability, going beyond traditional creativity support systems, such as video editing software, visualisation tools, or software development environments that Shneiderman (2007) lists as examples. In this sense, it can be considered to be both a computational creativity system and a creativity support system.

A fully functional beta version of the system was used in this evaluation. The same version is being used regularly by a small number of teachers and their pupils. More thorough evaluations of the educational use of the tool are being conducted separately from this study as a part of the on-going development effort. Instead of becoming a commercial product, the system aims to increase knowledge on human–computer co-creative systems and their pedagogical capability.

The main educational motivation for the Poetry Machine is to help pupils overcome creative droughts: When starting a new poem with the system, the pupil selects a topic from a list. The system then generates a five line long poem excerpt on that topic showing each word in a separate box simulating a fridge magnet. The pupil can edit the excerpt by moving, modifying, adding or deleting the words. The pupil can also ask for more materials in three ways: First, by dragging a magic wand over a word for replacement suggestions; second, with a rhyming tool that gives different types of rhymes for a word dropped on it; and third, by pushing a robot button, which generates more lines to the poem. The poetry writing interface of the Poetry Machine is shown in Figure 1.

The Poetry Machine is aimed at 9–12-year-old children. Thereby, the test participants in this case study must be children in this age group. Children as evaluators are different from adults in several ways: Many of their skills are still developing, including cognitive skills, such as concentration (Markopoulos & Bekker, 2003) and problem solving (Hourcade, 2008), and physical skills, such as motor and perception skills (Hourcade, 2008). Also the relationship between children participating in the test and the adult researchers administering the test affects the overall test success (Patel & Paulsen, 2002) and may cause bias in testing

(Read & MacFarlane, 2006). However, it is important to note that there are large individual differences between children, and task specific factors, such as good domain information, social support and good instructions, can substantially boost children's abilities in testing (Hourcade, 2008).

Testing with children requires special considerations so that it can be done in a comfortable, secure atmosphere in which children feel free to express their feelings about the application to be tested. For this case study, the school surroundings offer a safe and familiar environment for the evaluations. Evaluations done at school, however, require special considerations: In addition to requesting written consent from the pupils' parents, and verbal consent from the pupils themselves, a research permit from the school district is required in Finland.

3. Evaluation metrics and their background

Since the Poetry Machine can be considered both as a computational creativity system and a creativity support system, and the focus of the study was on the human participant's user experience, potential evaluation metrics were gathered from these three fields. The user experience and creativity support system articles were searched from the ACM Digital Library, whereas the search for computational creativity articles was focused on the International Conference on Computational Creativity (ICCC) and related references. The search on computational creativity was first focused on case studies, but they were either too focused on a specific domain or not appropriate for evaluating a creative use process. Therefore, more theoretical papers on computational creativity evaluation were searched for instead. Similar concerns directed the search in creativity support systems and user experience towards more general reviews.

Nine articles were selected as the basis for the evaluation metrics. In these papers, altogether 59 partly overlapping metrics were presented (see Table 1 for more details), but only a few of the metrics could be included in the study due to time limitations in the evaluation sessions, and the children's cognitive and social abilities. The selected metrics are: Fun, Enjoyment, Expressiveness, Outcome Satisfaction, Collaboration, Ease of writing and Ownership. Productivity, speed and efficiency oriented metrics were left out from the list, as they have received quite a lot of criticism if used in evaluating software intended for children (Hourcade, 2008) or creative contexts (Shneiderman, 2007).

The first metric, Fun, is a general user experience evaluation metric often used when evaluating children's software. It is considered a useful descriptor for user experience (Sim, Cassidy, & Read, 2013), but also important for keeping children's attention in educational software (MacFarlane, Sim, & Horton, 2005), and for assessing their willingness to use a product altogether (Read & MacFarlane, 2006; Sim, MacFarlane, & Read, 2006). Fun appeared as an evaluation metric in both the user experience evaluation review by Lachner et al. (2016) and in the creativity support system literature in the description of casual creators by Compton and Mateas (2015).

The next metric, Enjoyment, builds on the Fun metric. It derives from the work of Compton and Mateas (2015), more specifically, from their Pleasure parameter, but has also a strong connection to the Enjoyment metric described in the Creativity Support Index evaluation tool by Cherry and Latulipe (2014). Compared to Fun, Enjoyment clearly stretches over a longer time frame. As a measure of Enjoyment, Cherry and Latulipe ask their test

Table 1. Relevant background literature for the evaluation metrics as well as the final metrics they contributed to.

Field	Reference	Candidate metrics	Final metrics
UX	Lachner et al. (2016)	Appealing visual design; Communicated information structure; Visual branding; Mastery; Outcome satisfaction; Emotional attachment; Task effectiveness; Task efficiency; Stability and performance (9)	Fun; Enjoyment; Outcome satisfaction; Ease of writing (4)
CC	Ritchie (2001, 2007)	Quality, Novelty, Typicality (3)	Outcome satisfaction
	Colton, Charnley, and Pease, (2011)	Well-being rating; Cognitive-effort rating (2)	Outcome satisfaction
	Colton (2008)	Skillfull; Appreciative; Imaginative (3)	–
	Jordanous (2012)	Active involvement and persistence; Dealing with uncertainty; Domain competence; General intellect; Generation of results; Independence and freedom; Intention and emotional involvement; Originality; Progression and development; Social interaction and communication; Spontaneity/subconscious processing; Thinking and evaluation; Value; Variety, Divergence and experimentation (14)	Expressiveness; Outcome satisfaction; Collaboration (3)
	van der Velde, Wolf, Schmettow, and Nazareth, (2015)	Originality, Emotional value (of end result), Novelty/innovation, Intelligence, Skill (5)	–
CCS	Cherry and Latulipe (2014)	Exploration, Expressiveness, Immersion, Enjoyment, Results worth effort, Collaboration (6)	Enjoyment; Expressiveness; Outcome satisfaction; Collaboration (4)
	Resnick et al. (2005)	Support exploration; Cater for a variety of skill levels; Support many paths and many styles; Support collaboration; Support open interchange; Make it simple; Choose 'black-boxes' carefully; Design tools to be enjoyable; Use a multi-method approach; Use an iterative design approach; Design for designers; Evaluate your system (12)	Collaboration
	Compton and Mateas (2015)	Fun, Sense of ownership, Playfulness, Powerfulness, Pleasure (5)	Fun; Enjoyment; Ownership (3)

Notes: UX denotes the field of User Experience, CC the field of Computational Creativity, and CSS the field of Creativity Support Systems

users, if they would like to use the software regularly. This way, the users need to assess if the overall Enjoyment was impressive enough to motivate future use.

The Creativity Support Index by Cherry and Latulipe (2014) inspired also three other metrics: Expressiveness, Outcome satisfaction and Collaboration. Expressiveness describes how well the users are able to be creative and express themselves in the creative process. It is connected to Jordanous' (2012) metrics called Intention and Emotional involvement, which deal with self expression and emotional fulfilment in the creative tasks. Both metrics are presented in Jordanous' extensive study of creativity related phenomena in computational and human creativity literature.

The metric of Outcome satisfaction stems from the user experience evaluation review by Lachner et al. (2016). It implicitly requires the users to evaluate the final result of the creative process, so it is also related to the Quality, Value (Ritchie, 2001, 2007), and Well-Being effect (Colton et al., 2011) of the end result, used to rate artefacts produced by computational creativity systems.

The Collaboration metric is common in the creativity support literature. However, the component of creativity called Social interaction and communication that Jordanous (2012) presents in the computational creativity literature describes this metric more appropriately,

as it includes mutual influence, sharing and feedback with different agents - even without human collaborators. The Collaboration (Cherry & Latulipe, 2014) and Support collaboration (Resnick et al., 2005) in creativity support literature, on their part, usually require another person for collaboration.

Ease of writing is a component of Task Efficiency presented by Lachner et al. (2016). Efficiency as such is not a suitable metric for creative systems, so subjective efficiency was selected instead. This subjective metric is likely to reflect the overall user experience.

Finally, the Ownership metric is derived from the ideas of Compton and Mateas (2015). They emphasise a sense of ownership in user experience, and prioritise a feeling of control in the creative search process over finding objectively valuable artefacts in the search space. Ownership seems especially important when evaluating the use of co-creative systems.

Ten Likert scale statements and ten comparative questions were derived from these metrics, to be used in the questionnaires. Collaboration was considered an important topic for co-creative processes, and therefore it was evaluated with two questions. Expressiveness was similarly considered important for creative software and also evaluated with two questions. Ease of writing was evaluated for both the ease of starting the writing process and the ease of finishing the process. Fun and Enjoyment as related metrics were examined with one question each. Outcome satisfaction was considered clear to measure with one question, whereas Ownership was challenging to frame as multiple questions.

Many interesting metrics were left out of the list due to the context of the evaluations. For example, the Exploration metric, apparent in many studies, such as Jordanous (2012), Cherry and Latulipe (2014), and Resnick et al. (2005), was a very strong candidate. However, it was challenging to formulate questionnaire items that would fit both writing with the system and with a friend. Similarly, the Immersion metric would have been an intriguing metric, pointed out in several studies, such as those by Jordanous (2012), and Cherry and Latulipe (2014), and also by Compton and Mateas (2015), as they discuss the flow state. However, it was too difficult to fit it into the limited schedule, and it also seemed cognitively too challenging for the children. Some works were excluded altogether. For example, the work of van der Velde et al. (2015) did not reveal additional metrics considered applicable for this study.

4. Methods

The focus of the study were the experiences of the pupils writing poems with the Poetry Machine. The study looked at the aspects of the computational partner that promote a good co-creative experience, the pupils' reactions to computational creativity, and if they considered the program as an active collaborator or just as a tool. To get a wider notion of the roles pupils give and take in writing poems in collaboration, several conditions were compared, including human–computer (H–C), human–human–computer (H–H–C), and human–human (H–H) partnerships:

- H–C: One pupil and the Poetry Machine
- H–H–C: Two pupils and the Poetry Machine
- H–H: Two pupils on a paper

The condition of each pupil writing a poem alone on a paper would have been an interesting contrast. Unfortunately, it had to be left out because of the tight schedule of the test sessions, and the potential frustration of those pupils who are not used to writing poems. Having an entirely non-computational human–human interaction process as a comparative test condition enables the recognition of natural interactions between humans that may not be facilitated by the computational tool.

All pupils were asked to try out all the test conditions, because their personal working methods and habits would have otherwise made the results in different conditions incomparable. This within subject design exposed all participants to all test conditions increasing the probability of getting statistical differences between the test conditions with quite small sample sizes.

Twelve pupils participated in the evaluations. The participants were divided into three pairs of boys and three pairs of girls. Four of the participants were 10 years old at the time of testing, while eight were 11. Paired-user testing was used, because children are usually more relaxed in test situations if there are other children present (see, e.g. Höysniemi Hämäläinen, & Turkki, 2003).

The order of the conditions was randomised resulting in six different orders of test conditions, one test for each order. The participants of the study were recruited at school, and also the tests were conducted at the same school to provide a familiar and safe test environment.

Each test session lasted about 75 minutes and included 13 phases:

(1) Introduction
(2) Background questionnaire
(3) Test condition 1
(4) Post-task questionnaire 1
(5) Test condition 2
(6) Post-task questionnaire 2
(7) Break (with juice and biscuits)
(8) Test condition 3
(9) Post-task questionnaire 3
(10) Post-test questionnaire
(11) Post-test questionnaire walk-through
(12) Post-test interview (open questions)
(13) Thanks and stickers as reward

The first testing day started with a brief introduction in the pupils' classroom, introducing the researchers and emphasising the role of the pupils as experts giving valuable insight into how children interact with the system. Thereby, the introduction at the beginning of each session was very brief, and the test quickly moved on to filling in the background questionnaire asking about the pupils' age, gender, mother tongue, other languages spoken at home, things done with a computer, and interests in poem writing.

In each test condition, the pupils had only one test task: to write a poem. At first, the pupils were recommended to write a poem to congratulate a friend who likes animals to give an example on where to start, but this seemed needless and restrictive in most sessions, and the pupils usually started with a topic of their own, or one provided by the Poetry Machine. In test condition H–C, both the pupils received a laptop that they used to write

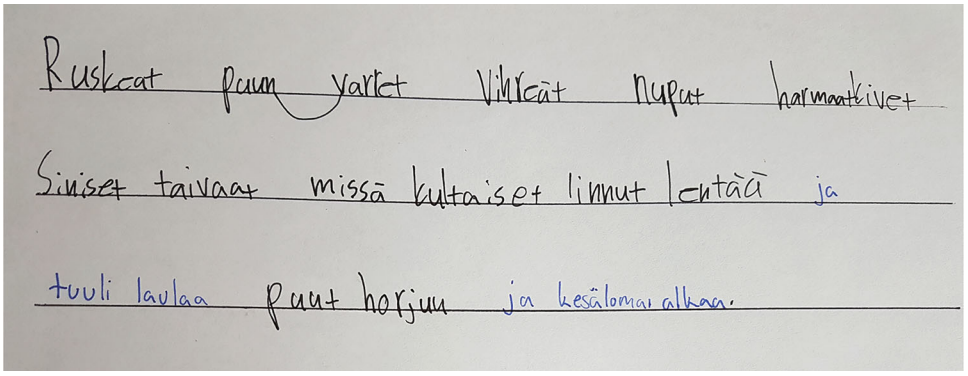

Figure 2. A poem written by one pair of pupils during testing, showing how the use of different colour pens captures the editing history.

their own poem with the help of the Poetry Machine. In test condition H–H–C, the pupils wrote a poem together with the help of the Poetry Machine running on a laptop. In test condition H–H, the pupils wrote a poem together on a paper using black and blue permanent pens. They were asked to cross words or lines out if they wanted to make changes to the poem, and to leave the original one still readable. An example of such a poem can be seen in Figure 2. Pen and paper were chosen over word processors, since 10–11-year-old children may not be familiar with a specific program and may get carried away by layout details and other visual aspects of the processors. Moreover, the different colour permanent pens produced a log comparable to the log produced by the Poetry Machine.

There were two researchers, both the authors, present in each test session to observe the pupils' reactions and interaction with each other, as well as their comments, tone of voice and gestures, and to give support in using the Poetry Machine when necessary. At times, the pupils required encouragement to get them started for example in finding a favourable topic. Sometimes they also needed to be reminded that they were not going to be graded or criticised for their poems, as they seemed to be worried about the quality of their poems.

All the test conditions were video recorded: in test conditions H–C and H–H–C, Flashback Express Recorder was used to record both the screen and the web camera feed. Also the log files from the Poetry Machine were available from these conditions. In test condition H–H, a separate video camera was used to shoot the writing on a paper. The same camera was also used to record the post-test interviews. An image of a live test session is shown in Figure 3.

After each test condition, each pupil filled in a post-task questionnaire that was the same for all the test conditions. It consisted of the 10 Likert statements provided in Table 2. A five-point rating scale based on smileys was used to give the ratings between completely disagree and completely agree. The pupils filled in their own copies of the questionnaires according to their own experience, but they were allowed to discuss the questions with their friend and also to ask the evaluators for clarifications. An example of a filled-in post-task questionnaire can be seen in Figure 4.

After all the three test conditions had been tried out, each pupil was asked to fill in a post-test questionnaire. It contained 10 comparative questions asking the pupils to rank the different writing methods. The questions are linked to the 10 Likert scale statements as shown in Table 2. To make it more fun and physically engaging to answer this last

Figure 3. One of the researchers observes as two participants write poems in condition H–C. The image shows the small classroom and the positioning of the equipment used in the study.

Table 2. Post-task and post-test questionnaire items by metric.

Metric	Post-task Likert statement		Post-test question	
Fun	$Q1_{ta}$	Writing the poem was fun	$Q1_{te}$	What was the most fun way to write a poem?
Enjoyment	$Q2_{ta}$	I would like to write poems in this way in the future	$Q2_{te}$	With what method would you prefer to write poems in the future?
Expressiveness	$Q3_{ta}$	I was able to be creative	$Q3_{te}$	With what method were you able to be the most creative?
	$Q4_{ta}$	I was able to express my own thoughts well	$Q4_{te}$	With what method were you able to best express yourself?
Outcome satisfaction	$Q5_{ta}$	I am happy with the poem I wrote	$Q7_{te}$	With which method did you write the poem you were most satisfied with?
Ease of Writing	$Q6_{ta}$	It was easy to start writing	$Q8_{te}$	With which method was it easiest to start writing a poem?
	$Q7_{ta}$	Writing was easy	$Q9_{te}$	With what method was it easiest to finish a poem?
Collaboration	$Q8_{ta}$	I got good ideas from the other writers	$Q5_{te}$	In which method were you able to get the best ideas from others?
	$Q10_{ta}$	Other writers were helpful for my writing	$Q6_{te}$	With which method were others the most supportive?
Ownership	$Q9_{ta}$	The finished poem is mine	$Q10_{te}$	With which method did you write the poem you felt was most your own?

Note: The subsrcipts *ta* and *te* denote post-task and post-test questionnaire items, respectively.

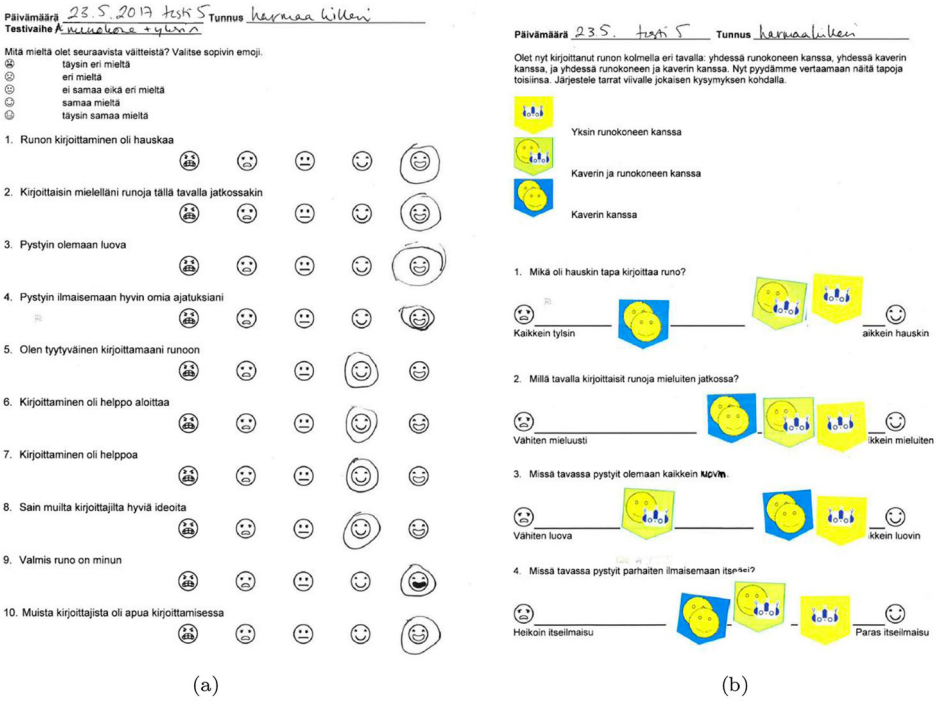

Figure 4. Examples of a filled-in post-task questionnaire (a) and the first page of a filled-in post-test questionnaire (b).

questionnaire, different coloured stickers with images symbolising the Poetry Machine (H–C), a friend and the Poetry Machine (H–H–C) and two friends (H–H) were designed to be used in comparing the different test conditions. A green sticker with a robot was used for H–C, a yellow sticker with a robot and a smiley face was used for H–H–C, and a blue sticker with two smiley faces was used for H–H. The pupils were asked to rate all the writing methods in all the questions. They were also allowed to rate the writing methods as equal, if they liked. An example of a filled-in post-test questionnaire can be seen in Figure 4.

Once the post-test questionnaires were filled in, they were used as the basis of a post-test questionnaire walk-through to give the children a chance to elaborate their answers. After the walk-through, a half-structured paired interview with a few open questions was conducted to examine further aspects related to the test conditions. A list of the open questions can be seen in Table 3. Both the walk-through and the half-structured interview were conducted so that both the pupils were present, and were usually asked to start answering in turns. At the end of the sessions, the children were thanked for and given a set of small stickers as a reward.

5. Analysis

The analysis began with a statistical analysis of the questionnaire results. A one-way analysis of variance (ANOVA) with repeated measures was conducted using the IBM SPSS statistics. The answers for each question were studied both by the condition (H–C, H–H–C or H–H)

Table 3. Themes for the half-structured interviews.

#	Question	Additional questions
1.	In your opinion, what aspects are important in writing poems?	Why?
2.	In which method did you focus best on writing the poem?	Why?
3.	In what way were you creative?	
4.	Was the Poetry Machine creative?	How?
5.	In which ways did you support others' work?	Did your pair notice your support?
6.	To whom would you recommend the Poetry Machine?	Why?
7.	Would you like it, if your school would start using the Poetry Machine?	Why, why not?
8.	What were the main benefits in using the Poetry Machine?	
9.	Was writing poems with the Poetry Machine similar to writing poems with a friend?	How?
10.	Did you learn anything about writing poetry, when working with the Poetry Machine?	What did you learn?
11.	Would you like to say something else about the Poetry Machine?	Would you change something about the Poetry Machine?
12.	Would you like to participate again in a similar test?	Why, why not?

and by the order of the conditions (1st, 2nd or 3rd). For this analysis, the post-task questionnaire results were encoded by awarding each answer one ("completely disagree") to five ("completely agree") points. For the post-test questionnaire, the first ranked method was awarded three points, the second two points, and the last one point. The Results section highlights all statistically relevant results, with significance level $\alpha \leq 0.05$.

A preliminary analysis of the video material collected only the most important comments from the pupils. A more thorough transcription described the features used in the Poetry Machine and the progress of the poems, as well as all the comments and questions made during the writing process.

The post-test walk-throughs and interviews were first transcribed, and then analysed by question. This allowed making connections to specific metrics, as well as connecting some of the questions to each other. After that, an analysis per interviewee was made comparing all the post-task and post-test questionnaire answers and the interview answers both within the interviewee and between their pair in the test. This helped in making a holistic picture of the pupils' experiences, and helped to identify potential problems in the writing processes.

The observations made in the test sessions were complemented with the analyses of the video material. Compared to the observations, the recordings gave a direct view to the pupils' faces enabling the interpretation of their facial expressions. However, there could be several explanations for frowning, yawning and laughing, so these gestures were not classified.

6. Evaluation results

This section presents the results for each of the chosen metrics: Fun, Enjoyment, Expressiveness, Outcome Satisfaction, Ease of Writing, Collaboration, and Ownership. The results of the ANOVA are presented first, separately for each questionnaire: For the post-task questionnaire filled in after each test condition, and the post-test questionnaire filled in after all three conditions were finished. The ANOVA results are complemented with the numerical results of the post-test questionnaire, interview comments from the pupils, and observations made in the test sessions.

6.1. Fun

Fun was measured with one question in both questionnaires ($Q1_{ta}$ and $Q1_{te}$). No statistically significant results are found in the ANOVA for either questionnaire. In the post-test questionnaire, six pupils ranked H–H–C as the most fun method, four favoured H–H, one raised H–H–C and H–H as equal, and only one favoured H–C. Those favouring H–H–C used arguments such as working together with a friend, working on a computer, or generally having fun inventing things with a friend. The pupils favouring H–H gave less reasons for their choices, as one pupil just liked to work with a friend, and another one said it was easier to write on a paper as they did not have enough experience with the Poetry Machine yet. The one pupil rating H–H–C and H–H as equal liked both writing together and writing on a computer. The one pupil rating H–C the best did not like to make compromises, and when writing alone with the Poetry Machine, they could use their own imagination and make their own choices.

In the open questions, the pupils mentioned fun twice as a general important factor for poetry writing, and also twice as a general factor of creativity. It was also a common reason for willingness in participating similar tests again (5 mentions), and for hoping the school to obtain the Poetry Machine in the future (3 mentions).

6.2. Enjoyment

The enjoyment metric was measured with $Q2_{ta}$ and $Q2_{te}$ asking about the preferred method for poem writing in the future. No statistically significant results are found for either questionnaire for this metric. In the post-test questionnaire, six pupils favoured method H–H, four H–H–C, one H–C, and one rated all the methods as equal. Although there were no statistically significant results yet, the means of both questionnaires indicate a preference for method H–H–C. In the walk-through, the pupils argued this choice with willingness to try the Poetry Machine again with the support of a friend, the help received both from the Poetry Machine and the friend, and the usefulness of the words given by the Poetry Machine. Method H–H was favoured for the pupils' own handwriting, the easiness of writing by hand, and one pupil considered the poem excerpt suggested by the Poetry Machine as a restriction. Method H–C was favoured for the feeling of writing on your own, although with the help of the computer. The pupil rating all the methods as equal considered choosing their own topic as the most important factor when assessing enjoyment.

Although the questionnaire results are inconclusive, all except for two pupils were keen on recommending the system to others in the interview: two to younger pupils, two to older people, and the rest to everyone regardless of their age. Nine pupils would have also liked to use the Poetry Machine at their school in the future, either in their own classes or with younger children; two pupils were uncertain; and only one specifically stated they would not have liked to use it themselves, but considered it suitable for younger children.

6.3. Expressiveness

Two questions measured expressiveness both in the post-task and the post-test questionnaires. The first question focused on supporting creativity ($Q3_{ta}$ and $Q3_{te}$), and the second on advancing self-expression ($Q4_{ta}$ and $Q4_{te}$). No statistically significant results were

obtained when the methods were compared, but the order seems to have an effect on subjective creativity, as there is a statistically significant increase between the first and the second method in the post-task questionnaires (ANOVA $p = 0.027$ and pairwise $p = 0.035$).

Eight pupils selected working alone with the Poetry Machine (H–C) as the most creative method in the post-test questionnaire. They argued this selection with the need to come up with their own ideas, and the lack of compromises. On the other hand, one of the two pupils favouring method H–H valued the fact that they did not have to work alone and regarded it as a main rationale for selecting this method. One pupil selected method H–H–C, and one pupil selected both H–H–C and H–H, as this pupil considered discussing their ideas with a friend as an important factor contributing to their own creativity.

When considering self-expression, seven pupils rated method H–H the best in the post-test questionnaire, four pupils favoured H–C, and only one pupil selected H–H–C. Method H–H was endorsed for being able to work with a similar friend and thereby reaching high levels of self-expression. Method H–C, on its part, was praised for the ability to use one's own imagination, making one's own choices, and not receiving critique from others. The pupil ranking H–H–C as the best did not explain their selection.

Expressiveness could be observed during the tests as delighted remarks like "I got it", when the pupils got a nice idea for their poem. The observations also support the good ranking of writing on paper when considering self-expressiveness, as quite many pupils seemed very enthusiastic about writing on paper, and got some surprising and bizarre ideas that sometimes caused strong creative disagreements. These contradictions, on their part, may indicate a version of self-expression.

The interviews revealed quite well what the pupils considered as creative. For example, one pupil linked creativity and self-expression to each other saying that

> In my opinion, to be creative, is to express myself, and to say, what comes to my mind, without altering it to please others, so that it's good just for me. At least. I am not saying that no-one else could like it. But anyway, it would be good also otherwise.

In addition, five pupils strongly related creativity to inventing an idea or words for the poem. Six pupils considered the Poetry Machine creative, dealing with creative tasks, but two pupils were strongly against the Poetry Machine as a creative entity. Due to time limitations, four pupils were not asked about the creativity of the Poetry Machine.

6.4. Outcome satisfaction

Outcome satisfaction was measured with one question in each questionnaire ($Q5_{ta}$ and $Q7_{te}$). This metric seems to be strongly affected by the order of the methods, as the last option scored the highest both in the post-task and the post-test questionnaires. The difference between the first and the last method is statistically significant in the post-task questionnaires (ANOVA $p = 0.001$ and pairwise significance between 1st and 3rd $p = 0.007$ and 2nd and 3rd $p = 0.014$). The relevance of the order of the methods is also supported by the interviews, as two pupils admitted that practice during the test may have affected their judgement of rating the last poem the best.

The poems written with method H–C were rated as the best six times in the post-test questionnaire, using arguments such as the use of their own ideas, and other self-expression and ownership factors. Five pupils favoured H–H, mainly due to the

amusingness of their poem. One pupil rated H–H–C as the best for the outcome, but did not give any reasons for this selection.

Outcome satisfaction could be observed during testing as willingness and even enthusiasm of the pupils to read out loud their own poems to their friend and the evaluators. The enthusiasm also seemed to increase toward finishing a poem, indicating a pleasing result. Perhaps the best indicator, however, was the persistent request of one pair to bring their cooperative poems at home with them. Some pupils also often spontaneously brought up details from their poems in the interviews and in the discussions with their friend. Even so, a number of pupils did not express their enthusiasm for the outcome while writing in any noticeable way.

6.5. Ease of writing

There were two questions about the ease of writing, separating the ease of starting to write ($Q6_{ta}$ and $Q8_{te}$) and finishing a poem ($Q7_{ta}$ $Q9_{te}$). Neither question produced statistically significant results in the questionnaires. Six pupils rated method H–H as the easiest to start with in the post-test questionnaire. One of these pupils mentioned that with the company of a friend, they did not feel pressed to start immediately, and two other pupils noted they got ideas much faster in this method. Four pupils selected method H–H–C as the easiest to start with, giving little rationale, except for one pupil suggesting that this method helped them to get ideas much faster than the other methods. Two pupils rated method H–C as the easiest to start with, one of them noting that the words prepared by the Poetry Machine helped them at the beginning, and the other noting they could start with the first idea that came to mind, indicating that collaboration with a friend requested some unwanted discussion over the idea.

Considering the ease of finishing a poem, five pupils rated method H–H as the best, one noting that they could easily divide the work with their friend, each taking turns to write, while two others indicated that the prepared words of the Poetry Machine were a distraction to them. Four pupils selected method H–C as the easiest to finish the poem with, one stating that they benefited from the prepared words. Three pupils selected method H–H–C, but did not give any explanations.

6.6. Collaboration

Collaboration was measured with two questions, one related to the quality of ideas from other writers ($Q8_{ta}$ and $Q5_{te}$), and the other related to the amount of support from other writers ($Q10_{ta}$ and $Q6_{te}$). The statistical analysis of the post-task questionnaire suggests that method H–H–C gave the most support from others, and H–C the least: H–H–C has the highest mean and H–C the lowest with a statistically significant difference (ANOVA $p = 0.030$ and pairwise $p = 0.013$). The post-test questionnaire produced even stronger results, suggesting that H–C falls behind the other two methods in both the quality of ideas and support from others. This is indicated by statistically significant pairwise comparisons between both H–C and H–H–C, and H–C and H–H (ANOVA for $Q5_{te}$ $p = 0.000$ and pairwise between H–C and H–H–C $p = 0.000$ and between H–C and H–H $p = 0.002$; ANOVA for $Q6_{te}$ $p = 0.001$ and pairwise between H–C and H–H–C $p = 0.000$ and between H–C and H–H $p = 0.016$).

Unfortunately, the difference between H–H–C and H–H is not yet statistically significant, although H–H falls behind H–H–C in both cases.

Eight pupils rated method H–H–C as the best for getting good quality ideas from others in the post-test questionnaire often referring to the materials prepared by the Poetry Machine. Two pupils notably stated that they selected method H–H–C as there were two other contributors, including the Poetry Machine, to help in ideation. Three pupils rated H–H as the best, but did not clearly state why. One pupil rated H–H–C and H–H as equal, stating they got ideas both from their friend and the Poetry Machine.

The arguments for selecting the method with the best support from others were not as clear. Seven pupils selected H–H–C as the best, but only one vague reason was given. Four pupils selected H–H, two of them explaining they liked the fact that they did not have to do everything by themselves in general, and also having spelling help from their friend. One pupil rated both H–H–C and H–H the best, but without any specific reasons.

In the interviews, some pupils mentioned group working as an important aspect of poem writing. When asked, if writing with the Poetry Machine was similar to writing with a friend, the pupils seemed unsure and found the comparison hard to make. Even so, 10 pupils considered writing with a friend and writing with the Poetry Machine as dissimilar: five of them specifically mentioned peer support as the most important difference between these methods, and some brought out the fact that with the Poetry Machine they did not need to negotiate about the poem. Two pupils regarded writing with the Poetry Machine as similar to writing with a friend, since both the friend and the Poetry Machine came up with words to use in the poem.

6.7. Ownership

Ownership was measured with one question in each questionnaire ($Q9_{ta}$ and $Q10_{te}$) asking the pupils if the finished poem felt their own. In the post-task questionnaire, method H–C was rated the highest and H–H–C the lowest, with a statistically significant difference(ANOVA $p = 0.007$ and pairwise between H–C and H–H–C $p = 0.033$).

In the walk-through of the post-test questionnaire, it became apparent that the pupils had felt that they were writing alone when they were using the Poetry Machine (H–C). Nine pupils rated the poem produced with this option to have the greatest ownership, and two of them specifically reasoning that they had worked alone. Three pupils selected method H–H, but gave little explanation to their choices.

7. Discussion

The lessons learned in this case study should be useful in designing similar studies for other co-creative processes and in developing new co-creative systems. The experiences and notes made in the study are presented starting from the issues related to the study context and the test procedure, moving on to the methods and metrics used, and finishing with general issues and future work.

7.1. Issues with study context and test procedure

There are several important aspects to take into account in designing a good evaluation, even more so, when real contexts and real users are involved. In this case study, the school

environment made it easier for the children to relax in the tests, and minimised the evaluators' need to look after the children outside the tests, as the school personnel took care of this. It also facilitated communication between the test participants in a positive way: The first pair participating in the tests did not yet know what to expect, clearly adding tension that only slowly faded till the end of the session. As the second pair arrived, they were notably more relaxed than the first ones, and the third pair was almost cheerful already when arriving. It is important to acknowledge this grapevine-effect in which the experiences of the prior participants affect the expectations of the next ones.

Stressing the general idea that the focus of the study is on the test participants' personal experiences and thoughts on improving the tested system seemed to reduce test tension, and create an atmosphere in which the participants are free to critique both the tested system and the test set-up. To check that the pupils have got on well in the test the interviews were finished with the questions "Would you like to participate again in a similar test?" and "Why?". All the pupils were willing to participate again, most saying it was fun, and some saying it was more fun than being in a normal lesson.

A relaxed atmosphere is especially important when testing creative applications. Conducting the same, or a similar creative task multiple times in a guided test session can be strenuous to any individual. The expectations of the participants affect the atmosphere a lot, but simple things during the test session can also alleviate stress. Having small talk with the participants before the test session is vital to create an open atmosphere. A small break was also held during each session, between the second and the final poetry writing task. During this break, the children could chat freely, and enjoy some juice and biscuits. This seemed to boost their energy levels, and to completely relax the atmosphere for the rest of the session.

7.2. Issues with evaluation methods

Creative tasks need to be rather open and cannot be instructed in such a detail as for example productivity oriented tasks. The initial idea was to use a similar recommendation for all writing methods: write a poem to congratulate a friend who likes dogs (or other animals). However, this seemed to restrict the pupils too much, and after the first test, the task was reformed into a suggestion that they could use or neglect. Having no official task seemed to unleash the imagination of some children, but some needed a sample to which they could revert to, and several pupils seemed to pick out the general theme of animals or friendship from the suggestion. This general preference for free form tasks was also backed up by a comment of one participant saying "I do like to write, because if I may choose the topic myself and so, it's quite fun, but if one needs to do some boring thing, it's not for me, even with a computer."

Observation of children was quite challenging, as some pupils were very talkative, whereas some remained quite silent even when working with their friend. Even so, comments like "I don't know what to do" or "I can't think of anything" were good indications of creative droughts or trouble with the system and Ease of Writing; happy remarks and laughter indicated Fun; smooth cooperation with the friend implied positive Collaboration; and enthusiasm to read out loud the finished poems was a strong indication of Outcome Satisfaction. However, some of the non-verbal communication conflicted with the pupils' questionnaire and interview answers, as well as with the evaluators' first impressions: At

times, what seemed to be very boring, was reported as fun, and what seemed like frustration turned out to be concentration, when the details of the writing process could be analysed from the test recordings. Furthermore, Enjoyment, Expressiveness, and Ownership, were almost impossible to observe, and some of the questionnaire results on these metrics were very unpredictable. This implies that the questionnaires and the walk-through of the post-test questionnaire were essential, especially for gathering the experiences of the silent pupils.

Observations were most useful when evaluating the Collaboration metric. Observation of actual collaboration is vital for understanding the ways children work with each other and creative systems. However, as the post-test questionnaire walk-through and the interview results indicate, collaboration still has some experiential aspects that are not covered by observation alone. Even so, it seems essential that designers of co-creative systems observe both the use of similar systems and the analogous co-creative tasks by human collaborators to best support the co-creative processes.

The questionnaires, overall, proved to be a central tool for evaluating almost all of the chosen metrics. Their importance was stressed especially in assessing Fun, Enjoyment and Outcome Satisfaction. With these metrics, it was also important to have both an immediate post-task questionnaire and a comparative post-test questionnaire. For example, based on the post-task questionnaire alone, all the writing methods were generally regarded as fun, but the post-test questionnaire results reveal the tentative preference for method H–H–C. On the other hand, with only the comparative results, a wrong assumption could be made that method H–C was not fun at all. Furthermore, the post-test questionnaire walk-through could be used to ensure that pupils had understood the questions and answered according to their own experience.

Some post-task questions seemed difficult for the pupils to interpret: For example, the Expressiveness related statement "I was able to be creative" was clearly difficult for the children, and the same difficulty could be seen in the open interview questions related to creativity. In addition, statements 8 and 9 caused confusion because of their wording. Statement $Q8_{ta}$, "I got good ideas from the other writers" stimulated several spontaneous questions, such as "Who others; there was no one else?" or "Can I get ideas also from an object?" referring to objects seen through the window. The answer to these questions was either "How do you feel about it?", or "Did you also get ideas from your friend, or the Poetry Machine?". Also statement $Q9_{ta}$, "The finished poem is mine", prompted several questions of "What does this mean?". The answer "Did it feel like your own poem?", usually clarified the issue.

Some pupils were rather silent during the post-test questionnaire walk-through, and one pupil explained that it was very easy to rank the different methods, but more difficult to explain their rationale. Also in the interview, some questions were more difficult for the pupils, especially when we asked for an elaboration. However, even difficult questions, such as "Was the Poetry Machine creative?" produced some strong opinions, although the pupils did not always know how to defend them. For example, one pupil answered this question with a definite "Yes", but when asked "How was the Poetry Machine creative?" they answered "I don't know . . . Or actually, I do know, but I just can't express it in words."

The analyses of the questionnaires confirmed that it is very important to analyse the results both by the writing methods and their order. Based on the preliminary results,

the post-task questionnaire seems more prone to bias from the order, as analysing it by order produced two statistically significant results, favouring the last writing method. The post-test questionnaire did not show as clear a bias from the order of the methods, and the results overall seem to be more varied. These results, however, demonstrate that the randomisation of the order of the test conditions is essential in comparative test settings like this.

Analysing the post-task and post-test questionnaires also by the interviewees and by the pupil pairs revealed interesting points about the similarities and differences in their answers. For example, the scale that the pupils used in the three post-task questionnaires was surprisingly wide compared to several other studies with children where the feedback is over positive (e.g. Obrist et al., 2009; Read & MacFarlane, 2006): three pairs used the whole scale 1–5; one pair 2–5 and 3–5; one pair 2–5 and 4–5; and only one pair just 4–5. In the post-task questionnaire, the pupils agreed very much with their pair about their experiences, as 163 out of the 180 pairs of answers were either the same or differed only with one point, 13 differed with 2–3 points, and only 4 were just the opposite within the pair. Furthermore, the metrics Fun, Enjoyment and Outcome Satisfaction produced similar answers within all the pairs, although the scales used inside these questions were as wide as from 2 points to 5. In the post-test questionnaire, there was more variety within the pupil pairs, as out of the 60 pairs of comparisons, 32 were exactly in the same order within the pairs, 18 differed by one place, and 10 by two places making the best for one pupil the worst for their pair. It appears the shared events caused children to report similar experiences, but individual writing preferences showed as clear differences in some questions.

7.3. Issues with evaluation metrics and results

Although the Fun metric was expected to be an important metric in this case study, it did not help to differentiate the writing methods statistically. However, there is a slight preference towards H–H–C in the means, maybe because it combines both working on a computer and working with a friend, which were separately named as fun factors by the pupils. The results for the Enjoyment metric are very similar to Fun. The post-test walk-through indicated that even though the metrics are related in the pupils' opinions, Enjoyment reflects on the long-term use, as intended.

The first Expressiveness statement, "I was able to be creative" produced statistically significant results only for the order of the writing methods. Similarly, the Outcome satisfaction increased till the end of the test making the pupils most satisfied with the last poem they wrote. This implies that both Creativity and Outcome satisfaction improve with practise.

The results for the second Expressiveness statement, evaluating Self Expression, are not statistically significant, but indicate that Self Expression was rated best with the H–H method and worst with the H–C method. This reflects the pupils' comments and the evaluators' observations that the poem excerpt and words given by the Poetry Machine sometimes constrained the pupils' own ideas. This Self Expression result is interesting when compared to the statistically significant Ownership result: Ownership was rated highest, whereas the Self Expression the lowest in the H–C method. It seems that high Ownership is therefore not dependent on high levels of Self Expression. All in all, Ownership seems a very multifaceted term: For some pupils it was clearly related to the number of participants

writing the poem; for some to the workload they put into the writing process; and for some to the source of the ideas for the content and the wordings.

Neither of the questions related to Ease of Writing produced statistically significant results, showing no clear preference to any method. Even the first statement: "It was easy to start writing" did not differentiate the methods, indicating that the problem of an empty paper, which originally motivated the development of the Poetry Machine, is not as relevant as expected. The questions measuring the ease of starting to write and ease of finishing the poem did generate dissimilar results, as expected, but the answers concerning the finishing are debatable, as the time seemed to run out for some pupils, and they may have been tempted to leave the poem as it was instead of actually finalising it. These unpolished poems may have affected also the answers related to Enjoyment, Outcome satisfaction, and Ownership. Even so, these questions seem more appropriate metrics than for example measuring the time it took to start writing or to finish a poem, as non-verbal metrics can easily lead astray with a creative task. Overall, the most useful data about the ease of writing are the comments some pupils made on their starting difficulties.

Both Collaboration metrics, the first focusing on the quality of ideas and the second on support from other participants, show a statistically significant increase between H–C and H–H–C, as well as between H–C and H–H. The mean of H–H–C is also higher than H–H, although this result is not statistically significant. As both the quality of ideas and the amount of support gradually increase from writing alone with the Poetry Machine, through writing with a friend on paper, to finally peaking at writing with the Poetry Machine and a friend, it seems that the ideas and support by the Poetry Machine are considered poor, but the Poetry Machine still has some capability in both.

7.4. Additional issues and future work

The writing instrument seems to play a surprisingly large role when evaluating poetry writing processes with children. Writing by hand was important for some pupils when rating the H–H method best for Enjoyment, but a nuisance for some pupils who enjoyed almost everything that they could do with a computer. Therefore it seems that the user experience with any writing process in general is affected by a large number of factors that were not investigated in this study.

This experiment used a multi-method approach, and it clearly demonstrated the benefits of such an approach: The quantitative tools, including the questionnaires, revealed what the current state of the system is and highlighted areas for further development, for example in supporting collaboration. The qualitative tools, including observation and interviews, on their part, gave ideas on how to improve the user experience. For example, supporting collaboration could be improved by introducing an iterative poetry writing process in which the human and the system take turns in writing new lines for the poem, as suggested by one of the pupils.

In order to further study the usefulness of the chosen metrics and to fully leverage the power of observing the co-creative processes in different conditions, the authors intend to conduct another full set of evaluations in a different school. This enables the comparison of different co-creative processes, as the children's creative working methods seem too diverse to make even preliminary interpretations based on this first sample alone.

8. Conclusions

This paper proposed a human perspective in evaluation of human–computer co-creative processes, and presented lessons learned in a case study using Poetry Machine as the computational creative partner. The chosen evaluation metrics, Fun, Enjoyment, Expressiveness, Outcome satisfaction, Ease of writing, Collaboration and Ownership, as well as the methods, observation in paired-user testing, questionnaires and interview, worked well in the school environment, chosen as the study context. The questionnaires worked well combined with the post-test questionnaire walk-through, during which the participants could clarify the criteria for their ranking. The strongest evidence through the questionnaires was produced for the Ownership and Collaboration metrics with statistically significant results even in this small-scale study. The results indicated that the Poetry Machine system was not generally considered as a co-author in the writing process, but still had some contribution. Some statements especially in the post-task questionnaire seem to be highly dependent on the order of methods, as for example in Outcome Satisfaction and Expressiveness the order of methods proved statistically relevant making the last methods subjectively the best when measured with these criteria.

The rich data provided in this case study should prove useful when improving the Poetry Machine system or developing similar systems in the future. The experiment presented here, accompanied with practical notes on creating an encouraging testing atmosphere and avoiding too restricting test tasks for creative activities, should be useful for any researcher attempting to evaluate co-creative experiences. The results on the chosen metrics, although preliminary, demonstrate which aspects might be useful for differentiating co-creative experiences between different systems in the future.

Yet further work is needed to establish the relevance of user experience evaluations in the field of computational creativity at large: We need to see a variety of different domains, experiment with different criteria, and formalise investigative tools, such as questionnaires. But most of all, we need to see in which domains and in what way the experience evaluation results can be utilised to develop more fluent human–computer co-creative processes.

Acknowledgments

The authors want to thank all pupils, who participated in the evaluations for their time and input, as well as their teachers for helping with the practical arrangements.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Academy of Finland under grant 276897 (CLiC).

References

Bown, O. (2014, June 10–13). *Empirically grounding the evaluation of creative systems: Incorporating interaction design*. The proceedings of the fifth international conference on computational creativity, pp. 112–119.

Bown, O. (2015, June 29–July 2). *Player responses to a live algorithm: Conceptualising computational creativity without recourse to human comparisons?*. Proceedings of the sixth international conference on computational creativity, Park City, Utah, pp. 126–133.

Cherry, E., & Latulipe, C. (2014, June). Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer–Human Interaction*, *21*(4), 21:1–21:25.

Colton, S. (2008). *Creativity versus the perception of creativity in computational systems*. AAAI spring symposium: Creative intelligent systems, Palo Alto, California. March 26-28, 2008 pp. 14–20.

Colton, S., Charnley, J. W., & Pease, A. (2011, April 27–29). *Computational creativity theory: The face and idea descriptive models*. Proceedings of the second international conference on computational creativity, Mexico City, Mexico, pp. 90–95.

Compton, K., & Mateas, M. (2015, June 29–July 2). *Casual creators*. Proceedings of the sixth international conference on computational creativity, Park City, Utah, pp. 228–235.

Davis, N., Hsiao, C.-P., Yashraj Singh, K., Li, L., & Magerko, B. (2016). *Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent*. Proceedings of the 21st international conference on intelligent user interfaces, New York, NY, ACM, pp. 196–207.

Hourcade, J. P. (2008, April). Interaction design and children. *Foundations and Trends in Human–Computer Interaction*, *1*(4), 277–392.

Höysniemi, J., Hämäläinen, P., & Turkki, L. (2003). Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers*, *15*(2), 203–225.

Jacob, M., & Magerko, B. (2015, June 29–July 2). *Interaction-based authoring for scalable co-creative agents*. Proceedings of the sixth international conference on computational creativity, Park City, Utah, pp. 236–243.

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, *4*(3), 246–279.

Kantosalo, A., Toivanen, J., & Toivonen, H. (2015, June 29–July 2). *Interaction evaluation for human–computer co-creativity: A case study*. Proceedings of the sixth international conference on computational creativity, Park City, Utah, pp. 276–283.

Lachner, F., Naegelein, P., Kowalski, R., Spann, M., & Butz, A. (2016). *Quantified ux: Towards a common organizational understanding of user experience*. Proceedings of the 9th nordic conference on human–computer interaction, New York, NY, ACM, pp. 56:1–56:10.

MacFarlane, S., Sim, G., & Horton, M. (2005). *Assessing usability and fun in educational software*. Proceedings of the 2005 conference on interaction design and children, New York, NY, ACM, pp. 103–109.

Markopoulos, P., & Bekker, M. (2003). On the assessment of usability testing methods for children. *Interacting with Computers*, *15*(2), 227–243.

Obrist, M., Igelsböck, J., Beck, E., Moser, C., Riegler, S., & Tscheligi, M. (2009). *Now you need to laugh!: Investigating fun in games with children*. Proceedings of the international conference on advances in computer enterntainment technology, ACM, pp. 81–88.

Patel, M., & Paulsen, C. A. (2002). *Strategies for recruiting children for usability tests*. Meeting of the usability professionals association: FL (June 2002). Retrieved from http://www.air.org/usability/publications/christinepaulsen/recruitingchildren.pdf

Read, J. C., & MacFarlane, S. (2006). *Using the fun toolkit and other survey methods to gather opinions in child computer interaction*. Proceedings of the 2006 conference on interaction design and children, New York, NY, ACM, pp. 81–88.

Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., & Eisenberg, M. (2005). Design principles for tools to support creative thinking. http://repository.cmu.edu/isr/816.

Ritchie, G. (2001). *Assessing creativity*. Proceedings of the AISB symposium on artificial intelligence and creativity in arts and science, York, England, pp. 3–11.

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, *17*(1), 67–99.

Shneiderman, B. (2007, December). Creativity support tools: Accelerating discovery and innovation. *Communications of the ACM*, *50*(12), 20–32.

Sim, G., Cassidy, B., & Read, J. C. (2013). *Understanding the fidelity effect when evaluating games with children*. Proceedings of the 12th international conference on interaction design and children, New York, NY, ACM, pp. 193–200.

Sim, G., MacFarlane, S., & Read, J. (2006). All work and no play: Measuring fun, usability, and learning in software for children. *Computers & Education*, *46*(3), 235–248.

van der Velde, F., Wolf, R. A., Schmettow, M., & Nazareth, D. S. (2015, June 29–July 2). *A semantic map for evaluating creativity*. Proceedings of the sixth international conference on computational creativity, Park City, Utah, pp. 94–101.

Yee-King, M., & d'Inverno, M. (2016 June 27–July 1). *Experience driven design of creative systems*. Proceedings of the seventh international conference on computational creativity, Paris, France, pp. 85–92.